Check for updates

ZSEQ: an interactive DNA sequence analysis program designed for microcomputers

MARTIN J. BISHOP

Department of Zoology, University of Cambridge, Downing Street, Cambridge CB2 3EJ, U.K.

ZSEQ is a self-documenting interactive DNA sequence analysis package designed for small computers based on Z80 microprocessors. The SZEQ philosophy is to provide basic operations which can be economically performed on small machines. ZSEQ is written in BCPL, a language which has been carefully designed to be near optimal for flexible text-processing applications (Richards & Whitby-Strevens, 1979).

ZSEQ provides counting, listing, translating, filing, splicing and pattern-finding facilities. ZSEQ does not provide searches for genes, secondary structure determinations nor comparisons of two or more sequences, activities which are best carried out on larger machines.

The only sequence length limitation on the programs is the value of the maximum integer held in one word of the BCPL implementation. (On the Z80 this is 32766.)

ZSEQ is designed to work on input files in EMBL format and the program should be used in conjunction with the *EMBL Nucleotide Sequence Data Library User Manual* ('EMBL manual') (Cameron *et al.*, 1983).

Each entry in the EMBL database corresponds to a single sequence, and an entry is structured so that it can be easily read by humans or machines. Each entry is composed of lines, the first two symbols in a line indicating its type. An entry is identified by its ID line which carries a unique eight-letter identifier. Information lines follow, particularly notable being the FT lines which locate the features of importance such as exons and introns. The start of the sequence itself is after the SQ line and the entry is terminated by the // line. Input files to ZSEQ may contain multiple ID entries, so that automatic processing of many sequences is made possible. The option is also provided to select individual ID entries from a file containing many sequences.

ZSEQ will also work on an input file containing a sequence alone, or will extract multiple sequences from input files in SEQ or GENBANK formats. It will not convert the GENBANK sequence positions into an EMBL features table.

ZSEQ always verifies the symbols present in sequences, and has facilities for alteration of unrecognized symbols to the set defined in Appendix B.2 of the EMBL Manual. The symbol # may be used to pad aligned sequences [note that the symbol - (hyphen) cannot be used for this purpose as it means any of A, C, G, or T].

The output from ZSEQ will be either sequences in EMBL format suitable for storage on disc for further processing or output suitable for printing.

ZSEQ is divided by function into four activities:

(1) *Counting*: counts may be made of bases, doublets, triplets, oligonucleotides of up to six in length, and codons.

(2) *Filing*: output may be sent to disk for further processing. Formatting, splicing, complementation, reverse complementation and translation may be performed.

(3) *Listing*: sequences may be listed with position numbers as single or double-stranded DNA or translated in up to six reading frames.

(4) *Pattern finding*: partial or total matches to an input pattern may be found in one or both strands.

When running in automatic mode, ZSEQ will process each ID entry in a file from position 1 to the // line. This allows rapid processing with minimal effort. In manual mode each individual ID entry may be selected, and then parts of the sequence may be further selected by position numbers. If a sequence is specified to be circular, reading over the physical end/beginning of the sequence may be performed.

Part of the additional power of ZSEQ over other analysis systems comes from its exploitation of the EMBL features table. Selecting the F option in manual mode enables the FT lines to be read by the program. The FT lines in the current ID entry are displayed by typing the symbol ? in response to the prompt for a key name. Feature table key names (Appendix A of the EMBL manual) may be selected to simplify specific tasks. For example, by selecting the key name CDS gene splicing may be performed with the introns discarded. This facility is available in all parts of the package, not merely for file output.

Starting

The package is intended to be self-documenting. In answer to prompts the following should usually have the desired effect:

? for help
control-Q to quit the current task
control-C to stop the program

Answers to prompts may be in upper or lower case letters.

Verifydna

This section checks for correct EMBL format of the input. If the input is not in EMBL format, the input is passed to EMBLFORMAT for conversion.

The sequence is then checked to ensure that it contains only the characters specified in Appendix B.2 of the EMBL manual. If there are characters in the input which do not conform to the standard but for which equivalents exist, the opportunity is given to convert them. A wild selection of characters at this point indicates that some arbitrary piece of text has been read in by mistake. Note that the Staden (Staden, 1979) or other uncertainty codes are not accepted by ZSEQ.

If the sequences are correct they are output for further processing.

Emblformat

This section converts sequences to EMBL format.

A sequence without other information is converted by the addition of an ID line, CC lines, FH line, FT lines, SQ line and // line.

Sequences is GENBANK or SEQ format are automatically converted to EMBL format so that they may be processed by the program. However, the GENBANK positional information is not converted to a feature table.

Identifiers in GENBANK or SEQ files may be up to 10 characters long. On conversion to EMBL format, identifiers have to be truncated to eight characters, but the words MADEFROM.⟨identifier⟩ are included on the ID line. Any clashes may therefore be distinguished even though they have to be acessed by the same search key.

Automatic mode

Automatic mode is used to process input files containing a large number of ID entries. Such files might typically contain homologous genes from a variety of organisms, or genes sliced from a single genome using the Filing option in manual mode operating on the EMBL FT lines. Terminal output is minimal in automatic mode, ID lines being reflected to keep track of processing.

Sequences are treated as linear molecules and are processed from position 1 to the // line of each ID entry.

Downloaded from https://portlandpress.com/biochemsoctrans/article-pdf/12/6/1015/493644/bst0121015.pdf by guest on 19 June 2020

Manual mode

Manual mode is used to process individual ID entries. The sequence may be accessed either by position numbers or from the FT lines.

An ID entry is accessed by giving its short identifier. This may be specified in lower case letters which will be converted to upper case by the program. Note that EMBL short identifiers must be in upper case. If a ? is typed instead of an identifier, a listing of the ID lines in the input file is displayed on the screen.

The DNA sequence may be treated as a circular molecule or a linear fragment. In the former case it is possible to read past the end of the sequence, as it appears on paper, into the beginning.

The format for giving sequence positions is 'From', 'To', 'Qualifier'. Normally 'From' is less than 'To' and the 'Qualifier' is blank. If 'From' is equal to 'To' a single base at that position is specified. If 'From' is downstream relative to 'To' and the qualifier is (C), the complementary strand is specified. These conventions follow the EMBL manual. When specifying sequence positions at the terminal the entries may be separated by spaces or commas. The abbreviations . (full stop) and * (asterisk) are available to represent the current cursor position and the end of the sequence respectively.

The input FT lines have to be in the fixed format specified by the EMBL manual. Otherwise the sequence positions are read in the same way as from the terminal. In addition an FT key name (Appendix A of the EMBL manual) may be selected. The key name may be typed in upper or lower case letters. Lower case will be converted to upper case by the program. Note that in EMBL files the key name has to be in upper case. To step through the FT lines '−' is typed at the terminal. Typing 'N' starts a new output entry and typing '+' adds results to the previous output.

Counting

Programs are available to count bases, to count overlapping doublets, to count overlapping triplets and to count non-overlapping triplets to give codon frequencies. In manual mode, typing '+' accumulates sums of sums until the next 'N' is typed.

When counting the base sum, the numbers of individual characters A, C, G, T, R, Y, N and # (Appendix B.2 of the EMBL manual) in the sequence are recorded as well as the numbers of purines (A+G+R), pyrimidines (C+T+Y) and all bases. U is treated as T, and the lower case letters as their upper case equivalents.

When counting doublets or triplets the deletion character # is ignored. If ambiguous characters (R, Y, N) are encountered, counts involving them are discarded.

When counting codons the deletion character is ignored, and any non-overlapping triplet involving ambiguous characters is discarded.

Oligonucleotides of lengths between four and six bases may also be counted, though the output may be voluminous.

Filing

The filing programs take sequences in EMBL format and output altered sequences in EMBL format to disc.

Selected parts of the sequences may be output. In addition complementation, inversion and translation operations are available.

This section makes it very straightforward to output the genes from longer sequences using the EMBL FT lines. For example, by selecting the key name TRNA all tRNA genes may be output. By selecting the key name CDS and using the 'N' and '+' controls, any type of RNA processing may be undertaken, with or without translation.

Note that if the qualifier (C) is set the output sequence fragment is complemented but not inverted.

The inversion operation is limited by the memory size of the machine. If this proves to be a problem, fragments may be inverted and the concatenated in the appropriate order.

Listing

DNA may be listed with position numbers as a single or double strand. Translation in all six reading frames may be added.

When DNA is listed without translation the standard output has 60 bases per line. This may be optionally altered to between 10 and 120 bases per line (in multiples of 10).

When DNA is listed with translation the line length is 45 bases giving 15 amino acids per line. Amino acids may be written in the three-letter or single-letter codes.

Pattern finding

Patterns of up to 1000 bases may be supplied. Patterns may contain ambiguous characters (N, R, Y). Matching is one sided: unambiguous bases in the pattern not being matched by ambiguous bases in the sequence. For example, Y in the search pattern will match T, C or Y in the sequence, whereas Y in the sequence will not be matched to T in the pattern.

In automatic mode a disk file of sequences in EMBL format is used to input the patterns. In manual mode each pattern is input at the terminal as required. Up to 80 bases may be typed, input being terminated by carriage return.

If a search for total matches is requested the output is in the form of position numbers. Matches in the complementary strand are marked by *.

Output from a search for partial matches displays the pattern and the portion of the sequence, with matches marked by colons.

Circular sequences have to be declared in manual mode in order that the physical end/beginning of the sequence may be searched. Negative position numbers mark the start of the match near the physical end of the sequence. Thus −2 is the position three bases before the end of the sequence.

Translation

Translation is available in both the Filing and Listing sections. The nuclear or mammalian mitochondrial codes are provided. In addition the user may alter either of these to the code table appropriate for other sequences. All that is necessary is to type the codons and the altered amino acid symbols which they represent.

Translated output may be in one-letter (Dayhoff) or three-letter codes. Codons involving ambiguous base symbols (R, Y, N) are always translated if possible, otherwise the symbol X (Xxx) is output. (Note that the aspartic and glutamic uncertainties are translated as X.)

When translating a gene with codons which are internally split by introns it is necessary to first output the DNA of the coding region and then to translate it. The translation programs do not automatically take care of this situation.

Implementation

ZSEQ has been implemented under a 64K CP/M operating system using the BCPL compiler described by Wilson & Webster (1980) and supplied by F. Fretwell-Downing Ltd. The executable code should run on any machine which has the CP/M operating system. The package is arranged with a main menu and seven overlay segments.

Though ZSEQ has been designed for a machine with limited memory, it will be easy to implement on any system

Downloaded from https://portlandpress.com/biochemsoctrans/article-pdf/12/6/1015/493644/bst0121015.pdf by guest on 19 June 2020

for which a BCPL compiler exists. It has been implemented unaltered, except for a small machine-dependent section, on an IBM 3081 computer under the MVS operating system. On this machine the BCPL word length is 32 bits, and ZSEQ will process DNA sequences of length up to 2147483647 bases. Because the sequence files are accessed by virtual I/O, also used on this machine to move arrays in and out of main store, the response is very fast.

Cameron, G., Hamm, G., Nial, J., Rudloff, A., Stoesser, G. & Stueber, K. (1983) *EMBL Nucleotide Sequence Data Library User Manual*, European Molecular Biology Laboratory, Heidelberg
Richards, M. & Whitby-Strevens, C. (1979) *BCPL – the Language and its Compiler*, Cambridge University Press, Cambridge
Staden, R. (1979) *Nucleic Acids Res.* **6**, 2601–2610
Wilson, I. D. & Webster, C. A. (1980) *Z80 BCPL System for CP/M and CDOS*, F. Fretwell-Downing Ltd., Sheffield

ZPEP: an interactive protein sequence analysis program designed for microcomputers

MARTIN J. BISHOP
Department of Zoology, University of Cambridge, Downing Street, Cambridge CB2 3EJ, U.K.

ZPEP is a self-documenting interactive protein sequence analysis package designed for small computers based on Z80 microprocessors.

ZPEP provides counting, listing, reverse translating, splicing and pattern-finding facilities. ZPEP does not provide secondary structure determinations nor comparisons of two or more sequences.

ZPEP is designed to work on input files in EMBL format but will also work on an input file containing a sequence alone, or will extract multiple sequences from input files in SEQ or GENBANK formats.

Input sequences should be in the one-letter code for amino acids. ZPEP always verifies the symbols present in sequences, and has facilities for alteration of unrecognized symbols to the set defined by IUPAC-IUB and given in Appendix B.4 of the *EMBL Nucleotide Sequence Data Library User Manual* ('EMBL manual') (Cameron *et al.*, 1983). The symbol // may be used to pad aligned sequences.

The output from ZPEP will be either sequences in EMBL format suitable for storage on disc for further processing or output suitable for printing. Output may be in the one-letter or the three-letter code.

ZPEP is divided by function into four activities:

(1) *Counting*: counts may be made of amino acids, doublets and triplets.

(2) *Filing*: output may be sent to disk for further processing. Formatting, splicing and reverse translation may be performed.

(3) *Listing*: sequences may be listed with position numbers in the one-letter or the three-letter code with reverse translation.

(4) *Pattern finding*: partial or total matches to an input pattern may be found in one or both strands.

Counting

Programs are available to count amino acids, to count overlapping doublets and to count overlapping triplets. In manual mode, typing '+' accumulates sums of sums until the next 'N' is typed.

When counting doublets or triplets the deletion character // is ignored. If ambiguous characters (B, Z, X) are encountered, counts involving them are discarded.

Filing

The filing programs take sequences in EMBL format and output altered sequences in EMBL format to disk.

Selected parts of the sequences may be output. In addition reverse translation is available.

Listing

Proteins may be listed with position numbers and reverse translation may be added.

When proteins are listed without translation the standard output has 60 amino acids per line. This may be optionally altered to between 10 and 120 amino acids per line (in multiples of 10).

When proteins are listed with translation the line length is 20 amino acids giving 60 nucleotides per line. Amino acids may be written in the three-letter or the single-letter code.

Pattern finding

Patterns of up to 100 amino acids may be supplied. Patterns may contain ambiguous characters (B, Z, X). Matching is one sided: unambiguous amino acids in the pattern not being matched by ambiguous amino acids in the sequence. For example, B in the search pattern will match D, N or B in the sequence, whereas B in the sequence will not be matched to D in the pattern.

In automatic mode a disk file of sequences in EMBL format is used to input the patterns. In manual mode each pattern is input at the terminal as required. Up to 80 amino acids may be typed, input being terminated by carriage return.

If a search for total matches is requested the output is in the form of position numbers.

Output from a search for partial matches displays the pattern and the portion of the sequence, with matches marked by colons.

Reverse translation

Reverse translation is available in both the Filing and Listing sections. The nuclear or mammalian mitochondrial codes are provided. In addition the user may alter either of these to the code table appropriate for other sequences. All that is necessary is to type the codons and the altered amino acid symbols which they represent. Reverse-translated output contains the Staden uncertainty codes (Staden, 1979):

R for A or G
Y for C or T
5 for A or C
6 for G or T
7 for A or T
8 for G or C
–for A or C or G or T (or any three out of the four.

Ambiguous amino acids (B, Z, X) are not translated but the symbols ??? are output.

Cameron, G., Hamm, G., Nial, J., Rudloff, A., Stoesser, G. & Stuber, K. (1983) *EMBL Nucleotide Sequence Data Library User Manual*, European Molecular Biology Laboratory, Heidelberg
Staden, R. (1979) *Nucleic Acids Res.* **6**, 2601–2610

Downloaded from https://portlandpress.com/biochemsoctrans/article-pdf/12/6/1015/493644/bst0121015.pdf by guest on 19 June 2020